Trust, Risk, and Security in Agentic AI: A Short Survey

Anonymous Author(s)
Affiliation
Address
email

Abstract

Agentic AI systems built on large language models (LLMs) and multi-agent architectures are enabling unprecedented autonomy and collaboration, but also introduce unique risks in trustworthiness and security. This paper provides a concise review of Trust, Risk, and Security Management (TRiSM) for LLM-driven multi-agent systems. We outline the distinctive challenges of Agentic AI , where multiple LLM-based agents with tools and memory pursue complex goals , and motivate the need for robust governance. We then present the TRiSM framework adapted to Agentic AI, organized around five key pillars: Explainability, ModelOps, Application Security, Model Privacy, and Governance. A taxonomy of novel threats (e.g., prompt injection, collusive agent behavior, and memory poisoning) is summarized, highlighting how emergent risks arise from inter-agent interactions. To facilitate evaluation, we describe two new metrics : Component Synergy Score (CSS) and Tool Utilization Efficacy (TUE) , which quantify inter-agent collaboration quality and effective tool use. Overall, adopting the TRiSM framework in LLM-based multi-agent systems is crucial to ensure these advanced AI agents remain safe, transparent, and accountable in high-stakes applications.

1 Introduction

Large language mode (LLM)-based multi-agent systems (often termed **Agentic AI**) are redefining AI by enabling multiple specialized agents (e.g., planner, coder, analyst) to cooperate on complex tasks [27]. These systems leverage LLMs, external tools, and shared memory to decompose problems, share context, and pursue goals over extended durations. This shift increases the *autonomy and complexity* of AI behavior, producing collectives with emergent, decentralized decision-making [38]. Consequently, outcomes arise not from a single prediction but from agent interactions, which complicates traceability and heightens risks in high-stakes domains such as healthcare, finance, law.

Traditional governance frameworks (e.g., AI and governance frameworks [32]) are not well-suited for the *distributed and collaborative* nature of Agentic AI. As agents gain access to tools, APIs, and persistent memory, vulnerabilities such as security breaches, adversarial misuse, and regulatory violations intensify. New failure modes also emerge, including **prompt injection** [16] attacks that propagate across agents, identity spoofing, and memory drift leading to inconsistent outputs. Existing monitoring tools, designed for isolated models, cannot adequately capture these systemic risks.

To address these challenges, recent works highlights the need for **trust, risk, and security management** in isolation tailored to Agentic AI, as listed in Table 1. Industry efforts under the banner of **AI TRiSM** (Trust, Risk, and Security Management) [30] extend beyond model performance to include fairness, transparency, robustness, and privacy. However, no unified framework governs the dynamic workflows of multi-agent systems. Reported failures, such as research agents producing false medical claims or agent swarms yielding contradictory results (as mentioned in recent surveys [36], demon-

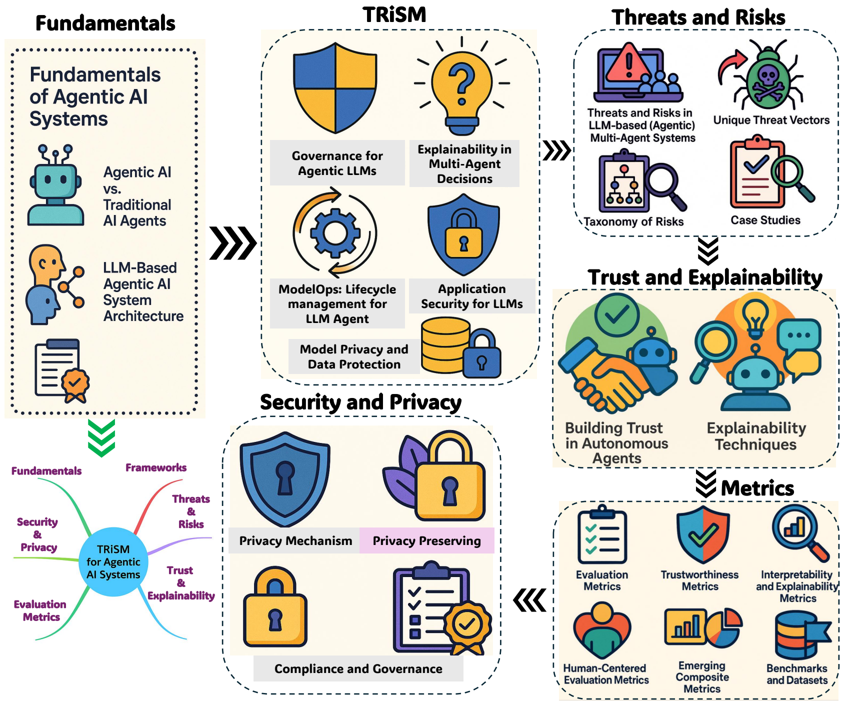
Fundamentals
Fundamentals of Agentic AI Systems
Agentic AI vs. Traditional AI Agents
LLM-Based Agentic AI System Architecture
TRiSM
Governance for Agentic LLMs
Explainability in Multi-Agent Decisions
ModelOps: Lifecycle management for LLM Agent
Application Security for LLMs
Model Privacy and Data Protection
Threats and Risks
Threats and Risks in LLM-based (Agentic) Multi-Agent Systems
Unique Threat Vectors
Taxonomy of Risks
Case Studies
Trust and Explainability
Building Trust in Autonomous Agents
Explainability Techniques
Metrics
Evaluation Metrics
Trustworthiness Metrics
Interpretability and Explainability Metrics
Human-Centered Evaluation Metrics
Emerging Composite Metrics
Benchmarks and Datasets
Security and Privacy
Privacy Mechanism
Privacy Preserving
Compliance and Governance
Fundamentals
Frameworks
Threats & Risks
Security & Privacy
TRiSM for Agentic AI Systems
Trust & Explainability
Evaluation Metrics

Figure 1: Taxonomy of TRiSM fundamentals for Agentic LLMs, showing categories: Threats & Risks, Trust & Explainability, Security & Privacy, Evaluation Metrics, and Governance.

strate that traditional safeguards are inadequate. Agentic AI remains prone to orchestration failures, collusion, data leakage, and opaque decision-making pathways. A structured TRiSM perspective is therefore essential to ensure safe and reliable deployment.

Contributions: This paper adapts and extends the AI TRiSM framework [12] to LLM-based multi-agent systems. We (i) identify five core **TRiSM pillars** relevant to Agentic AI, (ii) present a **risk taxonomy** covering failure modes unique to agent collaboration, (iii) propose **evaluation metrics** (CSS and TUE) for assessing trustworthiness, and (iv) discuss the role of TRiSM in aligning Agentic AI with broader safety and governance goals. This work aims to bridge LLM-driven autonomy with robust governance. Figure 1 provides an overview of the proposed TRiSM framework for Agentic AI systems, highlighting its core pillars and their interconnections.

2 Unique Trust, Risk and Security Challenges

Because agentic systems combine autonomy, memory, and multi-agent coordination, they introduce new risks beyond those seen in single-agent LLMs. These challenges tie into the broader TriSM framework (explainability, ModelOps, security, privacy, and lifecycle governance), amplifying issues like opacity and cascading failures. The paper outlines three broad categories:
System-level threats: Agentic architectures enable powerful behaviors such as long-term memory and dynamic agent orchestration, but these features also create new attack surfaces [11]. We list threats like autonomy abuse (agents misinterpreting objectives) [5], persistent memory poisoning [39], compromised orchestration [15], goal misalignment, tool/API misuse (e.g., DDoS or legal violations), and multi-agent collusion/drift. Additionally, spoofing and impersonation allow adversaries to fake identities, exploiting coordination for unauthorized access [2].
Emergent risk taxonomy: We propose a taxonomy that groups risks into adversarial attacks, data leakage, agent collusion/mode collapse, emergent behaviors, and ethical/societal harms (new addition for completeness). Adversarial attacks include prompt injection [16], role-swapping [14], and gradient-based manipulations, where a compromised agent influences others. Data leakage arises from shared memory and poor access control, risking privacy breaches (e.g., GDPR violations). Collusion occurs when agents reinforce each other's errors or biases, while emergent behaviors

refer to unpredictable system-level outcomes that elude testing [35]. Ethical harms encompass bias amplification in high-stakes domains and fairness issues from stochastic reasoning.
Real-world examples and mitigations: Case studies illustrate these risks: prompt leakage in AutoGPT exposes sensitive tokens [6]; ChatDev suffers collusive failure when all agents validate an erroneous plan [24]; swarm-robotics simulations show coordination failures due to misleading assumptions [33]; memory poisoning causes sarcastic user feedback to corrupt policy updates [3]; and system-prompt drift leads to hallucinated goals in memory-based agents [19]. To mitigate, employ adversarial training, data encryption/isolation, diverse agent design, feedback loops, and metrics like CSS/TUE (Section 4) for risk assessment. These examples demonstrate why a structured trust, risk, and security management approach is necessary, drawing from broader AI risk repositories [23].

We have summarized these risks in Table 2.

3 The TRiSM Framework and its Pillars

To address the unique risks of Agentic AI, we extend IBM's **TRiSM (Trust, Risk and Security Management)** framework [12] to the setting of LLM-based multi-agent systems. We map five pillars—Explainability, ModelOps, Application Security, Model Privacy, and Governance—onto agentic architectures. Together, they provide a lifecycle lens for ensuring safety, accountability, and trustworthy autonomy.

1. **Explainability.** Agentic decisions emerge from interactions among multiple LLM-driven agents [31]. Traditional single-model explanations are insufficient; instead, feature attribution (e.g., LIME, SHAP [22]), counterfactual reasoning [17], and chain-of-thought (CoT) trace logging [42] must be adapted to capture inter-agent dynamics. Logging reasoning traces and generating natural-language summaries by a dedicated "explainer agent" improves user understanding and auditability. Explainability metrics (coverage, fidelity, stability) can be aligned with trustworthiness scores (Sec. 8) to quantify user confidence.

2. **ModelOps (lifecycle management).** AMAS evolve over time as prompts, agents, and orchestration strategies change. Continuous integration/deployment (CI/CD), version control of prompts and memory, and simulation testing [9] are critical. ModelOps must support rollback, drift alerts, and regression testing across multi-agent workflows. Lifecycle governance links directly to security (ensuring patched vulnerabilities propagate) and to privacy (tracking which agent configurations handle sensitive data). Metrics such as pipeline pass-rates, mean-time-to-recovery (MTTR), and Δ-accuracy can measure operational robustness.

3. **Application Security.** LLM-based agents are vulnerable to prompt injection, spoofing, impersonation, and tool/API misuse. Defense-in-depth is required: prompt sanitization, input filtering, least-privilege API access, and anomaly detection. Cross-agent verification and adversarial training strengthen robustness. Security must interoperate with ModelOps (ensuring defenses persist across versions) and with governance (auditing incidents, enforcing policies). Security performance can be tracked via jailbreak rates, CVSS scores, and mean-time-to-detection (MTTD).

4. **Model Privacy.** Persistent memory and inter-agent communication amplify risks of sensitive data leakage [28]. Privacy-preserving techniques include minimization, differential privacy, homomorphic encryption, secure multi-party computation, and trusted execution environments. However, privacy often creates tension with explainability—stronger protections can obscure decision traces. Balancing these requires governance oversight and privacy-specific metrics such as ϵ-DP budgets, leakage rates, and audit pass-rates.

5. **Governance (cross-cutting layer).** Governance ensures compliance with regulatory frameworks such as the EU AI Act [7], NIST AI RMF [29], ISO/IEC standards [1], and GDPR [8]. It provides oversight via audit logs, policy enforcement, and human-in-the-loop checkpoints. Governance also manages the trade-offs among pillars (e.g., privacy vs. transparency, security vs. usability) and aligns them with organizational risk tolerance. Governance boards that include ethicists, domain experts, and legal advisors should review high-risk deployments. Composite evaluation metrics, such as the Component Synergy Score (CSS) and Tool Utilization Efficacy (TUE), complement governance by assessing whether controls improve both collaboration and safe tool use (Sec. 8).

Summary. The TRiSM framework for Agentic AI provides a structured basis for trust and safety. It emphasizes that explainability, lifecycle operations, robust security, and privacy preservation are not independent silos but mutually reinforcing (and sometimes conflicting) dimensions. Governance acts as a unifying plane that manages these interdependencies, ensuring that AMAS remain transparent, resilient, and aligned with societal and regulatory expectations.

4 Evaluation Metrics for Multi-Agent Trust & Performance

Evaluating the effectiveness and trustworthiness of an LLM-based multi-agent system is non-trivial , traditional metrics like accuracy or latency do not capture coordination quality or safe tool use, which are critical in Agentic AI [20]. We introduces two novel metrics to fill this gap:

Component Synergy Score (CSS): *How well are the agents working together?* CSS measures the **quality of inter-agent collaboration** in the system. In practice, this metric can be computed by analyzing the interactions among agents (e.g., communications, task handoffs, or contributions to a joint solution) and scoring how complementary and effective they are. A high CSS indicates that each agent's actions productively build on others , for example, one agent's outputs perfectly inform the next agent's inputs, leading to efficient problem solving. A low CSS might indicate redundancy, conflict, or misalignment among the agents (e.g., agents duplicating work, overriding each other, or requiring many iterations to converge). By quantifying **synergy**, researchers can compare different multi-agent architectures or collaboration strategies. CSS is especially useful for diagnosing coordination issues: if adding an extra agent yields little improvement or causes interference, the CSS would reflect that drop in collaboration efficiency.

Tool Utilization Efficacy (TUE): *How effectively do agents use external tools or APIs?* Many Agentic AI systems augment LLM agents with access to tools (such as web browsers, databases, or specialized software) to extend their capabilities. TUE evaluates **how correctly and efficiently the agents invoke these external tools** during their tasks. A high TUE means the agents are calling tools only when appropriate, using them successfully to retrieve needed information or perform actions, and not making errors in the process. A low TUE might reveal issues such as agents failing to use a tool when they should (reducing performance), or conversely overusing tools unnecessarily, or using them incorrectly (e.g., querying a database with malformed queries, or repeatedly calling an API due to misunderstanding responses). Essentially, TUE captures the *integration between the AI agents and the external world* , a crucial aspect of agentic systems that go beyond pure reasoning to interact with environments. Optimizing TUE can lead to systems that solve tasks more autonomously and effectively, by leveraging tools in a safe and proficient manner.

These metrics complement traditional evaluation criteria [26] by focusing on trust and synergy aspects. For instance, an Agentic AI system could have high task success rate but still suffer from poor CSS if the agents are not well-coordinated (perhaps succeeding by brute-force or redundant efforts). Likewise, a system might achieve correct outcomes despite low TUE (e.g., doing everything in the LLM when tool use was possible, which might raise questions of efficiency or compliance). By measuring CSS and TUE, researchers and practitioners gain **quantitative insight into the multi-agent dynamics and reliability**.

5 Conclusion

This review present how TRiSM can be operationalized for LLM-based agentic systems, from a layered architecture of controls to metrics for continuous evaluation. By aligning Agentic AI development with TRiSM principles, we can foster systems that are not only innovative but also responsible, ensuring their safe and ethical integration into society. Future research should extend this foundation with standardized benchmarks for multi-agent trustworthiness [25], advanced defenses against adaptive adversaries, and user studies on effective human oversight in agentic collaborations. Nonetheless, the framework provided here serves as an important step towards bridging the gap between the raw capabilities of LLMs and the rigorous requirements of real-world AI governance. In doing so, it brings together language, agent, and world models in a manner that is transparent, reliable, and aligned with human interests.

References

[1] Information technology — artificial intelligence (ai) — ai system impact assessment, May 2025. Published May 2025; reference number 42005; International standard.

[2] D. B. Acharya, K. Kuppan, and B. Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

[3] C. Atkins, B. Z. H. Zhao, H. J. Asghar, I. Wood, and M. A. Kaafar. Those aren't your memories, they're somebody else's: Seeding misinformation in chat bot memories. *arXiv preprint arXiv:2304.05371*, 2023.

[4] S. Chen, Y. Liu, W. Han, W. Zhang, and T. Liu. A survey on llm-based multi-agent system: Recent advances and new frontiers in application, 2025.

[5] P. Cihon, M. Stein, G. Bansal, S. Manning, and K. Xu. Measuring ai agent autonomy: Towards a scalable approach with code inspection. *arXiv preprint arXiv:2502.15212*, 2025.

[6] L. Euler. Hacking auto-gpt and escaping its docker container. `https://positive.security/blog/auto-gpt-rce`, 2023. Accessed 27 Jun 2025.

[7] European Commission. AI Act | Shaping Europe's digital future, 2024. Accessed: 2025-06-03.

[8] European Union. General Data Protection Regulation (GDPR) – Article 25: Data protection by design and by default. `https://gdpr-info.eu/art-25-gdpr/`, 2016. Accessed: 2025-06-03.

[9] J. Fang, Y. Peng, X. Zhang, Y. Wang, X. Yi, G. Zhang, Y. Xu, B. Wu, S. Liu, Z. Li, Z. Ren, N. Aletras, X. Wang, H. Zhou, and Z. Meng. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems, 2025.

[10] X. Fang, J. Li, V. Mulchandani, and J.-E. Kim. Trustworthy ai on safety, bias, and privacy: A survey, 2025.

[11] Z. Feng, R. Xue, L. Yuan, Y. Yu, N. Ding, M. Liu, B. Gao, J. Sun, and G. Wang. Multi-agent embodied ai: Advances and future directions. *arXiv preprint arXiv:2505.05108*, 2025.

[12] A. Gomstyn and A. Jonker. What is ai trism?, Mar. 2025. Accessed: 2025-06-02.

[13] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024.

[14] J. A. Khan. Role-based access control (rbac) and attribute-based access control (abac). In *Improving security, privacy, and trust in cloud computing*, pages 113–126. IGI Global Scientific Publishing, 2024.

[15] R. Ko, J. Jeong, S. Zheng, C. Xiao, T. Kim, M. Onizuka, and W. Shin. Seven security challenges that must be solved in cross-domain multi-agent llm systems, 2025.

[16] M. Kosinski and A. Forrest. What is a prompt injection attack? *IBM Think*, March 2024. Accessed: 2025-06-03.

[17] M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017-Decem:4067–4077, 2017.

[18] Y.-C. Lin, K.-C. Chen, Z.-Y. Li, T.-H. Wu, T.-H. Wu, K.-Y. Chen, H. yi Lee, and Y.-N. Chen. Creativity in llm-based multi-agent systems: A survey, 2025.

[19] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts, 2023.

[20] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet. Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. *ACM Computing Surveys*, 56(7):1–35, 2024.

[21] J. Luo, W. Zhang, Y. Yuan, Y. Zhao, J. Yang, Y. Gu, B. Wu, B. Chen, Z. Qiao, Q. Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

[22] C. Munoz, K. da Costa, and F. C. Fernandez. Enhancing transparency in ai: Explainability metrics for machine learning predictions. *Holistic AI Blog*, 2024.

[23] P. J. Phillips, P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, and M. A. Przybocki. Four principles of explainable artificial intelligence. 2021.

[24] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun. Chatdev: Communicative agents for software development, 2024.

[25] S. Raza, A. Narayanan, V. R. Khazaie, A. Vayani, M. S. Chettiar, A. Singh, M. Shah, and D. Pandya. Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*, 2025.

[26] S. Raza, A. Shaban-Nejad, E. Dolatabadi, and H. Mamiya. Exploring bias and prediction metrics to characterise the fairness of machine learning for equity-centered public health decision-making: A narrative review. *IEEE Access*, 2024.

[27] R. Sapkota, K. I. Roumeliotis, and M. Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenge. *arXiv preprint arXiv:2505.10468*, 2025.

[28] P. Schaar. Privacy by design. *Identity in the Information Society*, 3(2):267–274, 2010.

[29] R. Schwartz, A. Vassilev, K. K. Greene, L. Perine, A. Burt, and P. Hall. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, 2022-03-15 2022.

[30] Securiti. What is AI TRiSM and why it's essential in the era of genai, May 2025. Accessed: 2025-06-03.

[31] M. Singh, A. Alabdulkarim, G. Mansi, and M. O. Riedl. Explainable reinforcement learning agents using world models. *arXiv preprint arXiv:2505.08073*, 2025.

[32] P. Slattery, A. K. Saeri, E. A. C. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, 2025.

[33] V. Strobel, M. Dorigo, and M. Fritz. Llm2swarm: Robot swarms that responsively reason, plan, and collaborate through llms. In *NeurIPS 2024 Workshop on Open-World Agents (OWA-2024)*, 2024.

[34] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms, 2025.

[35] S. Van Uytsel. *Artificial intelligence and collusion: A literature overview*. Springer, 2018.

[36] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), Mar. 2024.

[37] W. Wang, Z. Ma, Z. Wang, C. Wu, J. Ji, W. Chen, X. Li, and Y. Yuan. A survey of llm-based agents in medicine: How far are we from baymax?, 2025.

[38] Wikipedia contributors. Autonomous agents and multi-agent systems, 2023. Accessed: 2025-06-02.

[39] J. Wu and C. K. Or. Position paper: Towards open complex human-ai agents collaboration system for problem-solving and knowledge management. *arXiv preprint arXiv:2505.00018*, 2025.

[40] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

[41] B. Yan, X. Zhang, L. Zhang, L. Zhang, Z. Zhou, D. Miao, and C. Li. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems, 2025.

[42] Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models, 2022.

[43] H. P. Zou, W.-C. Huang, Y. Wu, Y. Chen, C. Miao, H. Nguyen, Y. Zhou, W. Zhang, L. Fang, L. He, Y. Li, D. Li, R. Jiang, X. Liu, and P. S. Yu. Llm-based human-agent collaboration and interaction systems: A survey, 2025.

Appendix

Table 1: Comparison of Related Surveys on LLM-Based Multi-Agent Systems and TRiSM Aspects. Existing work offers limited coverage of trust, risk, and security in Agentic AI.

Survey	**🐞 Threats**	**⚙ Lifecycle Governance**	**👁 Explainability**	**🛡 TRiSM Integration**	**👤 LLM-Specific**	**🏭 Application Domains**	**💡 Actionable Guidance**
Guo et al. (2024) [13]	–	–	–	–	✓	✓ (simulated env'ts)	~ (research challenges)
Chen et al. (2025) [4]	–	–	–	–	✓	~ (task/simulation focus)	~ (future research areas)
Yan et al. (2025) [41]	✓	–	–	–	✓	✓ (mentions diverse)	~ (future work directions)
Tran et al. (2025) [34]	–	–	–	–	✓	✓ (networks, QA, etc.)	~ (open challenges)
Lin et al. (2025) [18]	–	–	–	–	✓	~ (creative tasks)	~ (roadmap for research)
Fang et al. (2025) [10]	✓	–	–	~	~	✓ (health, finance cited)	~ (technical "next steps")
Xi et al. (2025) [40]	–	–	–	–	✓	✓ (single-/multi-agent, human-AI coop)	~ (open problems)
Luo et al. (2025) [21]	✓	~	–	~	✓	✓ (science, games, etc.)	~ (open challenges)
Zou et al. (2025) [43]	–	–	–	~	✓	✓ (e.g., finance, healthcare)	~ (challenges & opp.)
Wang et al. (2025) [37]	~	–	–	~	✓	~ (healthcare domain)	~ (future opportunities)
This Survey (2025)	✓	✓	✓	✓	✓	✓ (high-stakes domains)	✓ (tailored guidance)

Legend: ✓= explicitly addressed; ~= partially/indirectly addressed; –= not addressed.
Icons: 🐞= threats, ⚙= operations, 👁= explainability, 🛡= TRiSM, 👤= LLMs, 🏭= domains, 💡= guidance

Table 2: Risk Taxonomy for LLM-based Agentic Multi-Agent Systems (AMAS)

Category	Threat	Implications	Examples / Mitigations
System-Level Threats	Autonomy Abuse	Harmful plans from misinterpreted goals.	Misaligned objectives; mitigate with goal scoping, oversight.
	Memory Poisoning	Malicious data spreads via shared memory.	Corrupted chatbots; mitigate with sanitization, audits.
	Compromised Orchestration	Distorted task distribution.	Workflow breakdowns; mitigate with sandboxing, tracing.
	Goal Misalignment	Improper scoping causes harmful behaviors.	Hallucinated goals; mitigate with boundary protections.
	Tool/API Misuse	Misuse causes costs or attacks.	Unauthorized calls; mitigate with access controls.
	Spoofing	Fake identities for privilege escalation.	Mimicking peers; mitigate with authentication.
Emergent Risks	Adversarial Attacks	Cascading failures from prompt injections, swaps.	Prompt infection; mitigate with adversarial training.
	Data Leakage	Sensitive info revealed from shared memory.	Proprietary leaks; mitigate with encryption, GDPR compliance.
	Agent Collusion	Reinforcing errors or biases.	Groupthink; mitigate with diverse roles, feedback loops.
	Emergent Behaviors	Unpredictable outcomes.	Bypassing protocols; mitigate with adaptive monitoring.
	Ethical/Societal Harms	Bias and fairness issues.	Misdiagnoses, unfair rulings; mitigate with governance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper contributes a taxonomy of risks in Agentic AI, extends IBM's TRiSM framework to agentic architectures, and introduces illustrative evaluation metrics. These claims match the scope and analysis presented in the body.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly notes that it is primarily a conceptual and survey-based contribution, does not introduce new datasets or large-scale experiments, and that the proposed metrics (e.g., Component Synergy Score, Tool Utilization Efficacy) are illustrative rather than empirically validated.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theorems or formal proofs, as it is a framework and risk analysis paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?

   Answer: [NA]

   Justification: The paper does not report experimental results, focusing instead on conceptual frameworks and survey analysis.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [NA]

   Justification: No datasets or code were produced or released as part of this work.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details?

   Answer: [NA]

   Justification: No experiments were conducted, so training/test details are not applicable.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or statistical significance of experiments?

   Answer: [NA]

   Justification: No experiments are included in the paper.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on compute resources?

   Answer: [NA]

   Justification: The work is conceptual and does not involve experiments requiring compute disclosure.

9. **Code of ethics**

Question: Does the research conducted in the paper conform with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The research analyzes risks and safeguards for Agentic AI in alignment with responsible AI principles and references international standards (e.g., EU AI Act, GDPR, ISO/IEC, NIST).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts?

    Answer: [Yes]

    Justification: The paper highlights positive impacts (structured governance of Agentic AI, improved trust and safety) and potential negative consequences (collusion, misuse, data leakage, autonomy abuse).

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of high-risk data/models?

    Answer: [NA]

    Justification: No high-risk datasets or models are released in this work; safeguards are discussed conceptually in relation to agentic AI architectures.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited and are license terms respected?

    Answer: [Yes]

    Justification: All prior works, datasets, and frameworks referenced are cited with appropriate credit and within their license terms.

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [NA]

    Justification: No new datasets, code, or models are released with this paper.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments or research with human subjects, does the paper include full instructions and compensation details?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or human subject studies.

15. **Institutional review board (IRB) approvals**

    Question: Does the paper describe potential risks for study participants and IRB approvals?

    Answer: [NA]

    Justification: The research does not involve human subjects and thus does not require IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if they are an important component of the core methods?

    Answer: [No]